The government seeks individual input; attendees/participants may provide individual advice only.

Middleware and Grid Interagency Coordination (MAGIC) Meeting Minutes[1]

March 6, 2019, 12-2 pm
NCO, 490 L'Enfant Plaza, Ste. 8001
Washington, D.C. 20024

Participants (*In-Person Participants)

Sachin Agarwal (USPS)	Brian Lin (UW-Madison)
Vinay Amatya (PNNL)	Miron Livny (UW-Madison)
V Anantharaj (ORNL)	David Martin (ANL)
Kathy Austin (TTU)	Shawn McKee (UMich)
Wes Bethel (LBL)	Ben Meekhof (UMich)
Laura Biven (DOE/SC)	Peter Nugent (LBL)
Ben Brown (DOE/SC)	Gilberto Pastorello (LBL)
Charlie Catlett (ANL)	Don Petravick (NCSA)
Dhruva Chakravorty (TAMU)	Ryan Prout (ORNL)
Vipin Chaudhary (NSF)	Hakizumwami Birali Runesah (UChicago)
Melissa Cragin (UI)	Sonia Sachs (DOE/SC)
Ewa Deelman (ISI)	Mat Selmeci (UW-Madison)
Sharon Broude Geva (UMich)	Sakshi ( )
Yong Chen (TTU)	Shalki Shrivastava (RENCI)
Florence Hudson (Northeast Big Data Hub)	Alan Sill (TTU)
Margaret Johnson (NCSA)	Jeffrey Tackes (USPS)
Joyce Lee (NCO)*	Jack Wells (ORNL)

Proceedings

This meeting was chaired by Vipin Chaudhary (NSF). February 2019 meeting minutes were approved.

Speaker Series: Data Life Cycle

- Peter Nugent, Senior Staff Scientist and Department Head for Computational Science, Lawrence Berkeley National Laboratory - *Multi-Messenger Astronomy and the Discovery of a Neutron Star - Neutron Star Merger*

[1] Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program.

- Charlie Catlett, Senior Computer Scientist, Argonne National Laboratory and Senior Fellow at the Mansueto Institute for Urban Innovation and Harris School of Public Policy, University of Chicago - *Using Computation and New Sources of Data to Understand Cities*

- Shawn McKee, Research Scientist, University of Michigan Physics Department, Director of the ATLAS Great Lakes Tier-2 Center, Director of the Center for Network and Storage-Enabled Collaborative Computational Science - *Scientific Data Lifecyle: Perspectives from an LHC Physicist*

Peter Nugent - *Multi-Messenger Astronomy and the Discovery of a Neutron Star - Neutron Star Merger*

First detection

Began in 1987, supernova in Large Magellanic Cloud observed. Prior to that, detected neutrinos at Kamiokande II. The neutrino Emission is by-product of the collapse of star. Photons do not get out until the shock wave of the super nova explosion, partly created by neutrinos, makes it way to the surface of the stars.

August 17, 2017: Diagram of detections at LIGO Hanford and LIGO Livingston and non-detection by Virgo gravitational wave detectors.

- Livingston: First detection of neutron star, neutron star (NS-NS) merger.
    - Structure starts at bottom left in middle panel and goes sharply at end when approaches 0 (zero) time – signal of 2neutron stars spiraling into each other and releasing gravitational wave energy.
    - Analysis and issuing an alert takes a few seconds,
    - 1.7 secs later – burst of gamma rays detected by the FERMI/GBM satellite
    - Hanford and Livingston– detection of spiraling, higher in frequency as get closer together. Merger (solid white line).
- Virgo – Did not detect as it was in a blind spot. As a result, could shrink error bars (dark green).
    - FERMI detection (GBM)- has uncertainty associated with gamma burst and interplanetary network honing into its location in the sky-- taken together, they led to a section of sky (see C). When we started to observe all galaxies at the right distance in the sky, we discovered a supernova-like event (part of NSF GROWTH project)
- Astronomical configuration of 3 things: random supernova, gamma ray burst and NS-NS merger

Global Effort

Follow up event: Observed everything possible for next 2 months (space, ground, gamma rays)
Are NS the long-sought sites of heavy element production?

UVOIR Light Curve spanning optical to infrared
Starts quickly, and visible in the ultraviolet, but quickly fades away as something very red. This event matched predictions made by scientists that heavy elements were synthesized.

Thumbprint of Heavy Elements
All the elements heavier than nickel in the periodic table. Model prediction of what NS-NS merger was to produce matched well with data of infrared at Gemini.

Periodic table of elements
Elements created in Big Bang; Once get to elements above 40, get to those produced by merging NS.

Q2. Jet Physics
Are NS mergers progenitors of short hard gamma-ray bursts (GRB)? Typically, think they are very energetic because it's pointed and directed at us. Event was weaker than a short GRB and the delayed onset of Radio/x-rays.

New Model: Cocoon Breakout
Cocoon for NS merger: Determined that there was a cocoon of material on the outside caused by the in-spiraling Jet explosion when merged and Jet rips through cocoon; GBR are immediately produced

Concordant picture of events: emitting gravitational waves; as gets closer, frequency increases; after Jet rips through material that was let off during the spiraling; after material bursts, get radio and x-ray delayed onset

Rates of these events – How common are these events
One event is possibly similar after vetting 150 candidates; upper rate of 1000/annually for 3 billion cubic light years; given this upper limit, this is consistent with NS-NS mergers being the main production sites of r-process elements in the Milky Way

After upgrades to the detectors, probably will find 3 of these events per year and 1 or 2 black hole/black hole mergers monthly.

Charlie Catlett – *Array of Things (AOT): A new Approach to Measuring Cities*
AOT–NSF- funded project partnership between ANL, UChicago, City of Chicago, Northwestern University
Project based on interactions with scientists:

- Partnering with City of Chicago to put up sensor network on light poles around city
- Find out what sensors that scientists wanted to put around the city. Some (transportation and social scientists) wanted to measure things for which there were no electronic sensors (e.g., analyze dangerous near misses)
- Needed remotely programmable computers inside these devices

- E.g., design safer street intersection to react in real-time, need to make decision on the street, no time for cloud. Programmable edge computing/software defined center
- AOT Configuration (FY18-19)/Nodes:
  - Environmental measurements (low cost, reliable)
  - Air quality (delicate)
  - Edge computing – use computer vision and audio analysis
- Mounted on street corners typically; not too reliable
- Data open and free via download, portals, real-time access; also, tools and tutorials
  - Data Consumer and Provider APIs – hooks for different languages
  - Sensor is a ML code: counting vehicles and pedestrians, every 30 seconds, which produces an observation
  - Framework based on ANL's open source platform (Waggle), which contemplates future pieces of hardware that will have new features or sensor; so API for registering device(s) to be added, including a network

<u>Sensors</u>

- 100 in place (blue dots) DOT is installing 10/week
- Have experimental interface; contemplated ways for residents to receive alerts of, for e.g., bad air quality, etc.; thinking of ways to make data more accessible to variety of populations
- Air quality –a couple of units at EPA site
- Tripomi satellite reading for most of gases measured daily; by 2020, NASA TEMPO satellite hourly with more granularity and of full atmosphere

<u>Science Examples</u>: lightening talks

- Illinois DOT funding to use ML software to obtain statistics on at-rate crossings
- DOE Vehicle Technologies Office project regarding transportation in and out of O'Hare Airport
- Also looking at how to use AOT image processing to detect flooding

<u>Exascale Computing-DOE project</u>

Looking at modeling the city from the point of view of energy performance and using AOT data to calibrate and check weather forecast models.

- E.g., Shanghai for every 1 degree Celsius that we can reduce the temperature of a city or district, results in 5 percent energy savings
- Chicago: how will buildings perform during extreme heat events?
- How to improve the fidelity of urban atmosphere models coupled with Energy plus (DOE software looking at energy performance of given building and their heat emissions)

Research partnership program (Green-shipped nodes; red: will ship soon)

Waggle Platform

- Edge computing, scaling, replication and being open
- Designed controller board to provide resilience to unreliable computers in the field
- Using open source hardware Linux platform; will be upgrading this summer
    - Performance – object recognition using Tensor flow (9 seconds computing time) vs. newer ML hardware (¼ of a second); so reports can be more aggressive in what can be analyzed

Community Engagement:
Worked with 7 schools. Trained over 400 students to build and deploy projects that would use wireless sensor technology.

Shawn McKee- ***Scientific Data Lifecycle: Perspectives from an LHC Physicist***
Challenges: due to growth, volume variety and velocity of data and corresponding impact on network and requirements.

Data Generation:

- Atlas-generation PB/second of data; not feasible to store. Process:
  1) Collect and select data (GBPS)
  2) Send data to a cluster of computers to be tagged by content and source
  3) Global data distribution-
- Problem in managing data and making it accessible
- Analysis: seeking new physics
- Visualization step – to understand data
- Interpretation – what does it tell us

Scale: 10s of PB of data (Slide 4)

- Access – through hierarchical "grid"
- LHC and experiments are upgraded; resources are evolving and being augmented with Cloud, HPC and new architectures

High Luminosity (HL)-LHC Challenge – Atlas and CMS needs in 2025, more complex data (Slide 5); need to address gaps (see diagram)

How are we addressing our challenges (Slide 6, examples)

- ATLAS Great Lakes Tier 2 (AGL T2) (Slide 7)
- Rucio - distributed data management service (Slide 8, link)
- OSiRIS - Open Storage Research infrastructure (Slide 9-10) - transparent high performance access to data, note Software Defined Storage Layer – exploring automation metadata that may be beneficial to science domains

- SLATE: Services Layer at the Edge (Slide 11) – scalable, multi-campus science platforms – centrally define subset of scientific applications
- OSG- Open Science Grid: common service and support for resource providers and scientific information
- IRIS-HEP- Institute for Research and Innovation in Software for High Energy Physics – focused on HL LHC (Slide 13)

Working Collaboratively (Slide 14)

Needed within and external to HEP community (e.g., ATLAS and CMS), including other science domains (astronomy/astrophysics) which are facing similar challenges

- Work together in the network
- Share workflow management systems

LHC Tool Trends and Technologies (Slide 15)

Important changes since LHC startup:

- Reduce arbitrary boundaries and definitions: migrated away from rigid infrastructure to more cloud-like resources which increases system efficiency
- Job workflow improvements:
    - Pilot job systems
    - Recent use of CERN virtual machines file system (CVMFS) and SQUID
        - Making centrally installed applications available onto distributed resources
        - looking at clouds and HPCS to facilitate running workflows over these resources
    - XCache – similar capabilities for large data sets
- Increasing network use – may use WAN access to date to alleviate storage requirements that would miss in HL-LHC
- Software refactoring – to take advantage of technology trends

Data Lakes (Slide 16)

Biggest challenge with HL-LHC is gap in data storage

- Storage is hard to manage and optimize across a large number of sites and too costly
- Create few large Data Lakes to provide needed service while reducing costs
- Status: under discussion and protyping; challenging to provide capability and capacity

Discussion

- Using edge devices to count cars, etc.; is any data being transmitted back or looking at high level entities (car, truck, etc.)? Pull information that want during analysis and send those integers back. Images get deleted due to privacy and cost issues.
- How validate algorithms? Every 15 min, pulling image from all the nodes, can use some images to train and evaluate models.
- Data available to public? Control access to training images via data use agreement.

Data Life Cycle Series Planning:

April - data use/re-use and data provenance/integrity, security and privacy

- Margaret Johnson and Don Petravick (NCSA) (Continuous Learning About Data: Experience from the DarK Energy Survey and NCSA) (confirmed)
- Yong Chen (TTU): Data provenance (confirmed)
- Victoria Stodden (UI) – reproducibility
- Lorena Barba (GWU) and Dan Katz (Journal of Open Source Software (JOSS)) data publishing

May: Data Triage: which data to discard and how the community decides what should be kept; perhaps exploring the process for deciding why some data can/needs to be discarded

- Dr. Jane Greenberg (Metadata Research Center; Drexel): big metadata (invited)
- Ben Blaiszik (Computation institute; UChicago); Material Data Facility http://materialsdatdafacility.org (proposed)
- Fran Berman (Rensselaer Polytechnic Institute and co-founder of RDA (RDA – developed working groups, interested in output)
- Digital Library Community
- Glenn Lockwood (NERSC) – large scale data storage

MAGIC Tasking (CY19)

Reports:

Containerization Series (Dhruva

DevOps Series (Alan Sill, contributor)

Anyone who is interested should contact Joyce Lee (joyce.lee@nitrd.gov)

Roundtable/Events

March 20-22, Coalition for Academic Scientific Computation (CASC) meeting, Alexandria, VA

- Academic Usage of Data Centers and Clouds ROI working group paper to come out soon

April 8 and November 3 deadlines, NSF Cyberinfrastructure for Sustained Scientific Innovation (CSSI) program solicitation

Next meeting

May 1 (12 pm ET)